GraphClust: A Method for Clustering Database of Graphs

Diego Reforgiato, Rodrigo Gutierrez, Dennis Shasha

Abstract

Any application that represents data as sets of graphs may benefit from the discovery of relationships among those graphs. To do this in an unsupervised fashion requires the ability to find graphs that are similar to one another. That is the purpose of GraphClust. The GraphClust algorithm proceeds in three phases, often building on other tools:

(1) it finds highly connected substructures in each graph;

(2) it uses those substructures to represent each graph as a feature vector; and

(3) it clusters these feature vectors using a standard distance measure. We validate the cluster quality by using the Silhouette method. In addition to clustering graphs, GraphClust uses SVD decomposition to find frequently co-occurring connected substructures. The main novelty of GraphClust compared to previous methods is that it is application-independent and scalable to many large graphs.